ARMY RESEARCH LABORATORY

Differential Effect of Correct Name Translation on Human and Automated Judgments of Translation Acceptability: A Pilot Study

Michelle Vanni and James Walrath

ARL-TR-4630 **September 2008**

NOTICES

Disclaimers

The findings in this report are not to be construed as an official Department of the Army position unless so designated by other authorized documents.

Citation of manufacturer's or trade names does not constitute an official endorsement or approval of the use thereof.

Destroy this report when it is no longer needed. Do not return it to the originator.

Army Research Laboratory
Adelphi, MD 20783-1197

ARL-TR-4630 **September 2008**

Differential Effect of Correct Name Translation on Human and Automated Judgments of Translation Acceptability: A Pilot Study

Michelle Vanni and James Walrath
Computational and Information Sciences Directorate, ARL

REPORT DOCUMENTATION PAGE

Form Approved
OMB No. 0704-0188

1. REPORT DATE *(DD-MM-YYYY)*	2. REPORT TYPE	3. DATES COVERED (From - To)
September 2008		

4. TITLE AND SUBTITLE
Differential Effect of Correct Name Translation on Human and Automated Judgments of Translation Acceptability: A Pilot Study

5a. CONTRACT NUMBER

5b. GRANT NUMBER

5c. PROGRAM ELEMENT NUMBER

6. AUTHOR(S)
Michelle Vanni and James Walrath

5d. PROJECT NUMBER

5e. TASK NUMBER

5f. WORK UNIT NUMBER

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)
U.S. Army Research Laboratory
ATTN: AMSRD-ARL-CI-IT
2800 Powder Mill Road
Adelphi, MD 20783-1197

8. PERFORMING ORGANIZATION REPORT NUMBER
ARL-TR-4630

9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)

10. SPONSOR/MONITOR'S ACRONYM(S)

11. SPONSOR/MONITOR'S REPORT NUMBER(S)

12. DISTRIBUTION/AVAILABILITY STATEMENT
Approved for public release; distribution is unlimited.

13. SUPPLEMENTARY NOTES

14. ABSTRACT

This study proffers two important findings: 1) automated machine translation (MT) evaluation is insensitive to the cognitive gravitas of proper names, contributing to its weak modeling of human judgments of higher quality MT output, and 2) there is a "new" methodology that produces superior measurement of translation acceptability. Twenty Arabic sentences, each with average name density of 3.7 names in 22 words, were translated into English with a research-grade MT system, to produce a 20-output-sentence Control Stimulus Set. Manual correction of 25% of the name translations resulted in an Enhanced Stimulus Set. A Magnitude Estimation (ME) methodology task had each of two teams of five subjects judge Control and Enhanced Sets against human reference translations. As is customary in ME studies, subjects made direct numerical estimations of the magnitude of the stimuli, in this case the degree to which sentences in the Sets conveyed the meaning in the reference sentences. Average estimates for Control and Enhanced Sets were 4.57 and 6.16, respectively, a 34.8% difference. Automated evaluation with the Metric for Evaluation of Translation with Explicit word ORdering (METEOR) produced scores of .446 and .546, a 22% difference. ME detected a differential effect, a finding which suggests that weighting proper name rendering in automated evaluation systems may improve correlations with human judgments on higher quality output.

15. SUBJECT TERMS
MT evaluation, name translation, magnitude estimation

16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT	18. NUMBER OF PAGES	19a. NAME OF RESPONSIBLE PERSON
a. REPORT	b. ABSTRACT	c. THIS PAGE			Michelle Vanni
U	U	U	UU	28	19b. TELEPHONE NUMBER (*Include area code*) 301-394-0367

Contents

1. Introduction **1**

2. Method **4**

2.1 Subjects 5

2.2 Apparatus....5

2.3 Procedure....6

3. Results **7**

4. Discussion **8**

5. Conclusions **9**

6. References **10**

Appendix A. **13**

Appendix B. **17**

Acronyms **21**

Distribution List **22**

INTENTIONALLY LEFT BLANK.

1. Introduction

The motivation for this study evolved from three related sources, the work of the Sequoyah Program Requirements and Data Collection Integrated Process Teams (IPT), feedback from Soldiers returning from Operation Iraqi Freedom and Operation Enduring Freedom, and the investigation of a fundamental applied research question.

Sequoyah is a Joint Army-led Acquisition Program of Record for foreign language translation. The program seeks primarily to equip military personnel with automated two-way speech and text translation capabilities in existing systems operating in mission-related environments and embedded at all echelons from a strategic level down to the individual war fighter. Its aim is to provide translation support for all languages on an annually updated language priority list. The level of linguistic support required is determined by the comprehensive mapping of level to tasks established by the Interagency Language Roundtable (ILR).

The "ILR level" is a designation of human foreign language proficiency which uses minimal semantic and syntactic correlates, and associated language-dependent tasks, as criteria. Within the USG language community, a literature has evolved in which tasks at each level are described. See, for example, Clifford and Lowe (1993). This paradigm considers completion of a specific foreign-language-dependent task as evidence of the learners' levels of either comprehension or production of either spoken or written language, depending on the task type.

For example, students who can read and translate editorials from foreign language newspapers and broadcasts (an ILR 3 task) would be considered "Level Three" in *written comprehension*. Students who can listen to and translate editorials from foreign language broadcasts (another ILR 3 task) would be considered "Level Three" in *spoken comprehension*. In spoken *proficiency*, however, learners must master tasks requiring spoken language production at a given ILR level in order to be considered proficient at that level. So, the learners who can invite you to a meeting (an ILR 2 task) are at least a "Level Two" in spoken production. But, unless they are linguistically capable of reasoning or expressing an opinion about the meeting or anything else (an ILR 3 task) they are not yet at "Level Three" in spoken proficiency.

In this context, a text machine translation (MT) evaluation research program has developed in which linguists seeking to perform tasks at specific ILR levels are assisted by the output of MT devices (Clifford, Granoien, Jones, Shen and Weinstein, 2004). A noteworthy by-product of this research has been discovering that a large number of MT systems are only minimally helpful to humans in performing at "Level One," the level where students are expected (among other things) to understand referring expressions and terminology, as with street names and menu items needed for reading maps and ordering at restaurants: ILR 1 tasks.

Realizing that military correlates to this level of proficiency abound in procedure manuals and training scenarios. The Sequoyah Requirements and Data Collection IPTs have expressed concern that Sequoyah MT systems handle language of complexities used in tasks at or near ILR 1. Thus, while it is expected that Sequoyah will augment existing MT system lexicons with military-specific terms, it is not often taken into account that names constitute an important noun subcategory for text and speech tasks performable at ILR 1.

Names, as tokens of this subcategory, are found interspersed among segments of data at higher levels as well. Names figure prominently not only in technologies designed for follow-on *automated* processing of MT output, but also in analytic *human* processing, understanding, or "sense making" of this material. In practice, MT systems are often imbedded within larger systems of multilingual information extraction and fusion. Many in the world of information technology envisage translation as the nexus of such a complex system, one that takes in foreign language images, speech, and text, and produces output which is English-like enough to support automatic functionalities such as entity and relation extraction, data fusion, summarization, and report generation, to name a few. Seen in this light, the challenge of proper name rendering in MT is fundamental.

Often disregarded in judging MT systems is the speed with which humans make sense of MT output. Moreover, the volume of correct content that humans glean from MT output is also rarely observed.[1] Yet these qualities reflect crucial characteristics of human interaction with MT output text, qualities that impact the types of information processing tasks that humans can perform with the output. Existing MT systems for speech and text reliably support only a minimum of information processing tasks, such as simple domain-specific dialogues and document triage.

A dialogue in a pre-specified domain between two monolinguals speaking different languages can be facilitated with two-way translation systems. Such MT implementations require that human processing of MT output take place in a timely manner. That is because the several tasks of formulating a reply, speaking or typing it into the machine, having the machine translate it, and allowing one's interlocutor to make sense of it, all need to occur within a Gricean space of time which signals cooperation.[2] In order for this to occur, naturally, systems have to be efficient. But, since human processing is also a factor, the output has to be understandable enough to support *human* efficiency as well. In this pilot experiment, we hoped to learn something about the extent to which name translation contributes to correct understanding of output by humans and, concomitantly, human efficiency in processing the output.

---

[1] This is the *Informativeness* metric of DARPA MT Evaluations in the 1990s (White et al., 1994).

[2] Paul Grice (1931-1988) was a British-educated language philosopher who formulated a description, known as the *cooperative principle*, of how people normally behave in linguistic interactions. From this, he derived a set of assumptions that listeners generally make about speakers in conversations. These are known as *Grice's Maxims*. See Grice (1975).

Document triage is generally performed by military intelligence analysts. When monolingual analysts are tasked to triage foreign language documents, MT systems equipped with domain-specific lexicons become valuable resources. Effective at the topic level, these systems' output can reliably contribute to the more complex task of human processing of *information* only when the elements of information, that is, specific extra-textual references are recognizable. Names are often among these elements. In this pilot, we sought to shed light on the extent to which appropriate name rendering in output contributed to accurate general understanding of output.

Appropriate treatment of the name subcategory may perhaps correlate with accurate and efficient human processing of MT output. If this is the case, then we, on the one hand, can encourage development in the direction of accurate name rendering and, on the other, expand the set of tasks to which we recommend the technology be applied. Translation contributes to correct human understanding of output that is beyond the topic level in order to extend the set of tasks to which we recommend the technology be applied.

Finally, our hypothesis that enhanced name rendering in MT can be effective in this way is based on two simple postulates in Applied Linguistics, a field in which, among other things, features of language material selected for exposition to learners is calibrated to the learners' level of proficiency.[3] Relevant concepts reside in the Input Hypothesis (IH) of Krashen (1981) and the Expectancy Hypothesis (EH) of Oller and Richard-Amato (1983).

Situations where readers attempt to make sense of MT output, and learners advance their knowledge of a second language by reading, are similar. In both, the interaction of language material with human cognition produces understanding at intermediate levels. A reader encounters a text, and an established level of language understanding is challenged. For learners of language, understanding is challenged by their level of progress in the learning process. But for MT readers, it is challenged by output quality.

In the IH, the term, *input* is used to refer to foreign language material to which a learner of that language is exposed, and the term, *intake,* refers to the sense made of that material by the learner. The IH advocates inclusion of "comprehensible input" for language learners. Similarly, in this pilot study, we probe improvement in comprehension when systems understandably render name mentions for MT readers.

The EH suggests that those exposed to a foreign language will use pragmatic or world knowledge to interpret the parts of the input that are beyond their levels of proficiency. This hypothesis is best understood in the author's own words:

> A person listening or reading, … is constantly generating hypotheses about what will come next in the sequence in terms of what the writer or speaker is intending to say. These hypotheses of the receiver are quite analogous to

[3]Here, we are not necessarily referring to ILR level.

> the plans of the sender. If the speaker's plans gibe closely with the hypotheses of the listener, communication is effective. If they fail to match well, communication breaks down. In both cases, the planning ahead or the hypothesizing what will come next can be conceptualized in terms of grammar-based expectancies (Oller, 1983).

New elements are interspersed with known elements for learners. Conversely, when names are understandably rendered in MT output, they constitute known elements, which get interspersed with "new" or possibly unclear elements that challenge the output reader. In the case of MT readers, the latter is unclear output, while in the case of language learners, the challenging material may consist of specific forms or structures that are beyond current levels of proficiency. In the MT case, system quality is the variable. In the learners' case, the learner's ability is the variable.

Learners' understanding is improved by providing challenging material along with known material. Thus, it follows that MT readers' understanding be improved by providing "known to be understandable" elements, such as interpretable names, among the often unordered renderings to which they are obliged to assign grammatical functions, modifications, and attachments in their attempts to understand the MT output. This is, in fact, our hypothesis. We want to observe the differential effect, if any, in levels of correct human understanding of output that is (1) a slightly revised raw version, and (2) a version that is edited to translate a defined quantity of names.

Automated metrics are the currently accepted yardsticks that we use to measure output quality. However, it is generally accepted that these programs are actually meta-metrics, which only approximate an actual measure. Current "non meta" measures involve human judgments of aspects of the output text in a comparison of a system (output) translation of the input with a reference human translation of the input. For this reason, we use both automated measures and human judgments of translation acceptability.

2. Method

Twenty Arabic segments (sentences) were translated into English segments using a research grade text-to-text MT system. These segments were drawn from open source material assembled in support of an annual MT competition.[4] The mean number of Arabic words per segment, was 16.8, and the mean number of English words per segment, in the resultant translation, was 26.

[4]The National Institute of Standards and Technology (NIST) annually conducts a competition among MT research systems. These data were part of the NIST 2008 Open MT Evaluation. For more information on this program, see http://www.nist.gov/speech/tests/mt/2008/doc.

Because the translations were to be judged by naïve subjects (i.e., non-linguists) they were modified to be somewhat more readable. For example, selections were made from alternatives (separated with slashes in the translation), and conjunctions such as "added follow" were changed to "added to." Descriptive and distance references were clarified, as were pronoun and prepositional references. Both the unmodified and the modified translation sets are presented in Appendix A. The modified translations formed the Control Stimulus Set.

The Control Set was then further modified. The number of correct name translations was manually increased by 25%. These 19 names were selected randomly (without replacement) from all the improperly translated names in the set. The names were then correctly translated. This segment set formed the Enhanced Stimulus Set and is also presented in appendix A.

The automated scoring of both the Control Stimulus Set and the Enhanced Stimulus Set was performed by the program named *Metric for Evaluation of Translation with Explicit ORdering* and known as METEOR (Lavie, et al., 2004). This metric is based on a generalized concept of unigram matching with ordering measured as a separate process. In METEOR, unigrams are matched based on surface forms, stemmed forms and meanings. It then combines unigram-precision, unigram-recall, and a measure that captures how well-ordered the matched words in the machine translation are in relation to the reference translation (Banerjee and Lavie 2005).

Since the subjects were monolingual English speakers, a reference translation of each of the 20 Arabic segments was also created by bilingual human translators. One-half of the subjects compared the segments from the Control Set with their respective reference translations, while the remaining subjects compared the segments from the Enhanced Set with the reference translations. The reference translations were identical to both groups and are presented in appendix A. Subjects were asked to judge the degree to which the machine translation conveyed the meaning present in the reference translation.

2.1 Subjects

Eight adult male and two adult female subjects volunteered for participation in this pilot study; they were non-linguists, and except for one, they were employed by the U.S. Department of Defense. No compensation was received for participation in the study, nor did any of the subjects have prior experience judging the acceptability of machine translations. Subjects were randomly assigned to one of two groups. The control group was presented with the Control Set of machine translations, while the experimental group saw the Enhanced Set of machine translations.

2.2 Apparatus

Subjects were given written instructions (appendix B) and a test booklet containing the 20 written machine translation segments appropriate for their group assignment. The instructions contained example translations that were not drawn from the National Institute of Standards and Technology (NIST) corpus, but rather were fabricated by the experimenters to assist in training.

Each page of the test booklet contained one machine translation, its reference translation, and a place for subjects to record their scores. (An example is presented in appendix B.)

2.3 Procedure

The experimental methodology consisted of a magnitude estimation (ME) task in which subjects compared machine translations (Arabic into English) with reference translations. Magnitude estimation is a method of psychophysical ratio scaling developed by S. S. Stevens in the early 1950's and has been frequently used in investigations as diverse as judging the brightness of a light, or the pitch of a tone to the prestige of occupations (Dawson and Brinker, 1971), or the goodness of moral judgments (Ekman, 1962). More to the point, ME methodology has also been used to measure linguistic acceptability (Bard and Robertson, 1996). Subjects in ME studies are asked to make direct numerical estimations of the magnitude of stimuli. In this experiment, the stimulus magnitude was defined as the meaning present in the reference translation that was also present in the machine translation (i.e., how much meaning survived the machine translation). Because ME yields ratio scale data, the full gambit of statistical testing can be used to analyze the data—in contrast to the ordinal data produced by Likert-like scaling techniques. A more extensive discussion of ME methodology is available from Grescheider (1985).

For this study, subjects were asked to consider how much of the meaning present in the reference translation was also present in the machine translation. They scored this as a number between zero and 10. Integers, fractions, or decimal numbers were allowed. Subjects were instructed that a score of zero meant the machine translation provided no hint of the meaning expressed in the reference translation. A score of 10 would reflect a machine translation that perfectly conveys all of the meaning in the reference translation (neither adding nor losing any meaning).

At the end of the written instructions an example translation and reference translation were provided. This example was identical to all subjects. Subjects were asked to score this example (see instructions to subjects in appendix B). The example was specially fabricated by the experimenters to represent a midpoint (score of 5) on the zero to 10 scale. Subjects were then asked to keep in mind the score they gave this example as they scored each of the 20 test segments. Thus, the example segment became what is called the modulus, or standard stimulus, against which all other segments were judged. For example, if subjects scored the modulus as a five, and felt that one of the test segments was only half as good at conveying the meaning present in the reference translation, they should assign a score of 2½ to the test segment.

A single independent variable (IV) was manipulated. This IV was the number of correct name translations present in the machine translated text segments scored by the subjects. Two levels of the IV were employed. They were the number of correct name translations occurring in the Control and Enhanced Stimulus Sets, as previously described. The dependent variable was the subject's ME score.

Keeping in mind that the only difference between the Control Stimulus Set and the Enhanced Stimulus Set was the number of names correctly translated, it was hypothesized that the Enhanced Stimulus Set would result in ME scores that were significantly greater than those scores obtained from the subjects judging the Control Stimulus Set. Further, it was hypothesized that the percent increase in ME scores (Enhanced Set compared to Control Set) would be greater than the percent increase yielded by the automated scoring methods. Confirmation of this hypothesis would indicate a differential effect of correct name translation on human and automated judgments of translation acceptability.

3. Results

The grand mean ME score for the modulus was 5.53.

The mean ME scores for the 20 Control Set segments and the 20 Enhanced Set segments were calculated. The grand means for the Control and Enhanced Sets were 4.57 and 6.16, respectively (a 34.8% difference). A t-test found a statistically significant difference between the groups, t= -2.685 with 38 degrees of freedom (P=.011).

Six of the 20 segments were identical in both the Control and Enhanced Sets (recall that only 25% of the incorrectly translated names in the Control set were changed to correct translations). No statistical difference, between the Control and Enhanced group's ME scores, was found for these six segments.

It is sometimes recommended that analysis of ME data be based on geometric rather than arithmetic means (Gescheider, 1985). The geometric mean is equal to the antilog of the sum of the log of the scores, divided by the number of scores. Geometric means were calculated for the 20 Control Set segments and the 20 Enhanced Set segments. The grand means for the Control and Enhanced Sets were 4.25 and 5.97, respectively (a 40.5% difference). A t-test found a statistically significant difference between the groups, t = –2.747 with 38 degrees of freedom (P = .009).

METEOR scores for the 20 Control Set segments and the 20 Enhanced Set segments were obtained. The grand means for the Control and Enhanced Sets were .446 and .546, respectively (a 22% difference). A t-test found a statistically significant difference between the groups, t = –2.36 with 38 degrees of freedom (P = .023).

METEOR scores were regressed onto the ME scores using an incremental order polynomial regression (through the third order). None of the resulting regression equations yielded a significant finding.

4. Discussion

Ultimately, the objective of this line of research is to examine the question of differential effects of correct name translation on human and automated judgments of translation acceptability. We did not know, a priori, whether a difference existed or, if it did, the magnitude of the differential. Therefore, it was necessary to implement a measurement tool that was very sensitive to differences in human judgments of translation acceptability. We felt that ME could provide this level of sensitivity. Thus, this pilot study was really two-fold in purpose: first, to explore the effects of correct name translation on human and automated judgments of translation acceptability, and second, to evaluate ME as a measurement tool.

Our results strongly suggest that there exists a differential effect and that ME is sensitive enough to measure it. Even the small sample size of five subjects per condition, yielded evidence that increasing the number of correctly translated names had a differential effect on human, versus automated scores (i.e., that a 25% increase in the number of correctly translated names resulted in an increase of nearly 35% in humans scores as opposed to a 22% increase in automated scores).

It is also interesting to note that on the six segments that were identical to both groups, the between group ME scores were not significantly different. This would seem to indicate that the significant difference in ME scores, between groups, on the entire 20 segment sets, was due in fact to the improved names and not to subject differences.

Further evidence that ME methodology is well suited to the measurement of translation acceptability comes from another study by us and will be published at a later date. This second experiment exactly replicated the methods and procedures of the current study with a single exception. Rather than have subjects rate the translation acceptability using ME, they used a traditional four-point Likert scale. Where the ME methodology produced data showing a highly significant difference in human judgments of translation acceptability between the Control and Enhanced sets of translations ($P = .011$), the data from this second study failed to show a difference. These results lend support to the argument that ME methodology provides a finer scale with which to measure translation acceptability.

That said, a careful review of the study suggests that two aspects of the employed methodology should be changed in future work. First, a score of zero should not be permitted as the logarithm of zero is undefined, and it is often useful to convert ME scores to logarithms. Second, there should be no upper limit placed on ME scores. Using an upper limit of 10, as we did in this pilot study, creates an unwanted ceiling effect in the data.

5. Conclusions

There is a general lack of consensus concerning the importance of correctly translating names. Understanding what causes the disagreement is crucial before automated systems can be made to better model human judgments. To what extent external factors influence the importance of correct name translation also warrants a closer look. For example, the 20 segments used in this study were selected from a known corpus for their name density, which averaged 3.7 out of 22 words per sentence. It is possible that longer segments with a smaller ratio of names to words would yield different results. In such studies, however, it would be important to factor in the effects of co-reference or the lack thereof.

An entity may not be identifiable when it is first introduced into the output discourse because the system has not properly rendered its name. In such cases, it is improbable that a reader of the output will correctly identify anaphoric pronominal references to that entity as co-references. Thus, name translation becomes important for attenuating the compounding of "sense making" problems in longer segments.

How might diverse features of a set of test segments influence the differential effect seen here? These characteristics might include the presence of: (1) names and co-references, (2) names and specific given information, (3) names and new information about the entities they represent, or (4) a sparser distribution of names. How might the context of use of an MT system influence human translation acceptability? How would an automated metric reflect this? Further investigation is necessary to answer these important questions.

6. References

Bard, Ellen G.; Robertson, Dan. Magnitude Estimation of Linguistic Acceptability. *Language* **1996**, *72* (1), 32–68.

Child, James R.; Clifford, Ray T.; Lowe, Jr., Pardee. Proficiency and Performance in Language Testing. *Applied Language Learning* **1993**, *4*.

Clifford, R.; Granoien, N.; Jones, D. A.; Shen, W.; Weinstein, C. J. The Effect of Text Difficulty on Machine Translation Performance -- A Pilot Study with ILR-Rated Texts in Spanish, Farsi, Arabic, Russian and Korean. *In Proc. 4th International Conference on Language Resources and Evaluation* in Lisbon, Portugal, ELRA, 24–30 May 2004.

Dawson, William E.; Brinker. Validation of Ratio Scales of Opinion by Multimodal Matching. *Perception and Psychophysics* **1971**, *9*, 413–419.

Doddington, George. Automatic Evaluation of Machine Translation Quality Using N-Gram Concurrence Statistics. *In Proceedings of 2nd Human Language Technologies Conference (HLT-02)*, San Diego, CA. pp. 128–132, 2002.

Ekman, Gösta. Measurement of Moral Judgment: A Comparison of Scaling Methods. *Perceptual and Motor Skills* **1962**, *15*, 3–9.

Grescheider, George A. *Psychophysics Method, Theory, and Application*, (2nd ed.) Hillsdale, New Jersey: Lawrence Erlbaum Associates, 1985.

Grice, Paul. *Logic and Conversation*; In *Syntax and Semantics, 3: Speech Acts*, ed. P. Cole & J. Morgan. New York: Academic Press, 1975. Reprinted in *Studies in the Way of Words*, ed. H. P. Grice, pp. 22–40. Cambridge, MA: Harvard University Press, 1989.

Interagency Language Roundtable Website (2006). ILR Language Skill Level Descriptions. http://www.govtilr.org.

Krashen, Stephen D. *Principles and Practice in Second Language Acquisition*; English Language Teaching series, London: Prentice-Hall International (UK) Ltd, 1981.

Banerjee, S.; Lavie, A. METEOR: An Automatic Metric for MT Evaluation with Improved Correlation with Human Judgments. *In Proceedings of Workshop on Intrinsic and Extrinsic Evaluation Measures for MT and/or Summarization at the 43rd Annual Meeting of the Association of Computational Linguistics (ACL-2005)*, Ann Arbor, MI, June 2005.

Lavie, A.; Sagae, K.; Jayaraman, S. The Significance of Recall in Automatic Metrics for MT Evaluation. *In Proceedings of the 6th Conference of the Association for Machine Translation in the Americas (AMTA-2004)*, Washington, DC, September 2004.

Melamed, I. Dan Automatic Evaluation and Uniform Filter Cascades for Inducing N-best Translation Lexicons. *In Proceedings of the Third Workshop on Very Large Corpora (WVLC3)*, Boston, MA, 1995.

Oller, John W.; Richard-Amato, Patricia A. Eds. *Methods that Work*; Rowley, MA: Newbury House, 1983, (pp. 4–5).

Papineni, Kishore; Roukos, Salim; Ward, Todd; Zhu, Wei-Jing. BLEU: a Method for Automatic Evaluation of Machine Translation. *In Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics (ACL-02)*, Philadelphia, PA, July 2002. (pp. 311–318).

White, J. S.; O'Connell, T. A.; O'Mara, F. E. Advanced Research Projects Agency Machine Translation Program: 3Q94. *Proceedings of the November 1994 Meeting*, 1994.

INTENTIONALLY LEFT BLANK.

Appendix A.

Appendix A contains the twenty original machine translations, the same translations modified to be somewhat more readable (Control Stimulus Set), the translations after manually increasing the number of correct name translations by 25% (Enhanced Stimulus Set), and the reference translations.

Table A-1. MT output, reference translations, control and enhanced stimulus sets.

Machine Translations	Control Stimulus Set	Enhanced Stimulus Set	Reference Translations
liar added follow the forces A for the no FGhA intention there QW following A JNBIH TTMRKZ imagining in/on pricked NH.	Addition to forces A for the no FGhA there are also foreign forces central station in pricked NH.	Addition to forces Afghan there are also foreign forces central station in Ghazni Province.	In addition to the Afghani forces there are also foreign forces that are headquartered in Ghazni.
sperms declared morning LHIIEH A ran realize land treads intention (fable fable SI) that source in Russia informed matter the assassination.	A.m. declared he to land treads intention (fable fable SI) that source in Russia informed him on matter the assassination.	A.m. declared he to British Broadcasting Corporation (BBC) that source in Russia informed him on matter the assassination.	In the morning he told the British Broadcasting Cooperation that a source in Russia had informed him of the assassination order.
attributed follow liar insulted A LSHAFH escape forgot saying for the no A support the withdrawal immediate from Irak or reduction provision forces.	To him was attributed A LSHAFH as saying for the not support the withdrawal immediate from Irak or reduction forces.	To him was attributed A LSHAFH as saying for the not support the withdrawal immediate from Irak or reduction forces.	He was quoted by the Agence France Press as saying: “I do not support the immediate withdrawal from Iraq or the reduction in funding our forces.”
liar considers logic sweeping rocks which falls about 60 spoke south with Baghdad who important forts the dustmen stands within what names triangle allegiance.	Region of logic sweeping rocks which falls about 60 km south with Baghdad important forts for dustmen stands within what names triangle.	Region of Jurf al-Sakhr which falls about 60 km south with Baghdad important forts for Al-Queda stands within what names triangle.	The area of Jaraf Al Sakher, which is located around 60 km south of Bagdad is one of the most important Al Qaeda's hideouts and is located within what is called the triangle of death.
clarified OBD praiseworthy that council security SIOQD JLSH next thursday listening follow making/report on/from visit A bite Eh follow number who countries A fled IQIH and from among it the Sudan.	Clarified OBD said that council security SIOQD JLSH next thursday listening a report on visit members to a number of countries A IQIH and from among it the Sudan.	Abdul Mahmood said that council security SIOQD JLSH next thursday listening a report on visit members to a number of countries African and from among it the Sudan.	Abdul Mahmood explained that the Security Council will conduct a session next Thursday to listen to a report on its members' visit to a number of African countries, including Sudan.

Table A-1. MT output, reference translations, control and enhanced stimulus sets (continued).

Machine Translations	Control Stimulus Set	Enhanced Stimulus Set	Reference Translations
came lip the advertisement/declaration after hours who indictment/accusation with Perez liar FSKI, during MWTMR journalist in/on London, the Russian President Vladimir Putin optimists/that matter personal with numbers Moa bitter its/his assassination.	Came the declaration after hours with Perez liar FSKI, during MWTMR journalist in London, the Russian President Vladimir Putin that matter his assassination.	Came the declaration after hours with Perez liar FSKI, during MWTMR journalist in London, the Russian President Vladimir Putin that matter his assassination.	This announcement came hours after Berezovsky accused, in a press conference in London, President Vladimir Putin of personally ordering the preparation of a conspiracy to assassinate him.
after weakened height JIH in Paris, returned night first who yesterday follow with Beirut, active in/on "the current the national free" mark Ho IK which right/afflicted in/on events 23 that the past	After weakened period in Paris, returned night yesterday back with Beirut, who active in "the current the national free" mark Ho IK hurt in events 23 january.	After weakened period in Paris, returned night yesterday back with Beirut, who active in "the current the national free" Mark Howiak hurt in events 23 january.	After a period of treatment in Paris, the activist of the Free Patriotic Movement Mark Howaik who was injured during the events of last January 23, returned to Beirut the night before last.
wants FIZ provokes Washington who Tehran liar or lapse T IKLF respects man annulment with program nuclear.	Wants FIZ provokes Washington from Tehran and or lapse designates respects man to annul program nuclear.	Chavez provokes Washington from Tehran and or lapse designates respects man to annul program nuclear.	Chavez Provokes Washington from Tehran and Olmert Delegates Lieberman to Thwart its Nuclear Program.
liar confirmed passer-by figs ga pulls, the spokesman the foreign ministry A for the no what intention only that in the citizens the german lost that in/on Afghanistan	Confirmed passer-by figs ga pulls, the spokesman the foreign ministry A only that in the citizens the german lost two in Afghanistan.	Confirmed Martin Jager, the spokesman the foreign ministry A only that in the citizens the german lost two in Afghanistan.	Martin Jager, spokesperson for the German Foreign Ministry, said that two Germans are missing in Afghanistan.
kidnapped MSLHW student N the Korean southern Ali in the way of MQATtOH pricked NH after the appearance Thursday.	Kidnapped armed student N the Korean southern on the path of MQATtOH pricked NH P.m. Thursday.	Kidnapped armed Taliban the Korean southern on the path of Ghazni Province P.m. Thursday.	Taliban gunmen kidnapped the South Koreans on a road in the Ghazni province on Thursday afternoon.
quotation fighters who detachments the brigades in/on incursion my creates/zionist in/on house for the no hurry.	Quotation fighters from detachments brigades in incursion a zionist in Beit Lahia death.	Quotation fighters from detachments brigades in incursion a zionist in Beit Lahia death.	Two Fighters from the Al Qassam Brigades Were Martyred in a Zionist Incursion in Beit Lahia.
Fatah THBTt its/his attempt enthusiasm holding the councillor middle hurling with trespass the constitution Israel kills 4 and for the Palestinian.	Fatah frustrates attempt enthusiasm holding the contract in trespass the constitution.	Fatah frustrates attempt enthusiasm holding the contract in trespass the constitution.	Fatah Thwarts Attempt by Hamas for "Legislative" Contract Amid Accusations of Violating Constitution.

Table A-1. MT output, reference translations, control and enhanced stimulus sets (continued).

Machine Translations	Control Stimulus Set	Enhanced Stimulus Set	Reference Translations
bases what KW now A love the descendant who Iosaka (west) year 1965 LJNH "the peace who polish Vietnam" or guided/according name my Japanese "BIHA rings".	Bases what came from the descendant Iosaka (west) year 1965 founded "the peace who polish Vietnam" or name Japanese "BIHA rings."	Bases what came from the descendant Osaka (West) year 1965 founded "the peace who polish Vietnam" or name Japanese "BIHA rings".	Oda who came from Osaka (West) founded in 1965 a "Peace for Vietnam" committee or "Beheiren" in Japanese.
Jerusalem brigades TFJR Abo H explosive in/on doctor door SHIW intention twist IH A supporter Salah Aldin border site Nahal the enrolment.	Jerusalem brigades detonate explosive in doctor barrel SHIW intention twist IH A supporter Salah Aldin bombs site Nahal the enrolment.	Jerusalem brigades detonate explosive in doctor barrel SHIW intention twist IH A supporter Salah Aldin bombs site Nahal the enrolment.	Al-Quds Brigades Detonate Explosive Device in Zionist Tank; al-Nasser Salah al-Din Brigade Bombs Military Site of Nahal Oz.
Warning T ro TINIH delay political America warns subjects from travel for the entity the Zionist my lands his/its palestinian/annulment.	Warning routine political warning America warns subjects from travel for the Zionist and my lands Palestinian.	Warning routine political warning America warns subjects from travel for the Zionist and my lands Palestinian.	Routine Warnings, Political Warnings. America warns its citizens against traveling to the Zionist entity and the Palestinian territory.
happened LLMNZzMH the English IH its/his arab that participated success in/on A divorce journalists French knives did arrested A in/on the Iraq.	Happened before LLMNZzMH the English IH arab participated success in A divorce journalists French arrested A in the Iraq.	Happened before the English-Arab Organization participated success in A divorce journalists French arrested A in the Iraq.	Earlier, the English-Arab Organization contributed successfully to the release of French journalists who had been detained in Iraq.
praised leader A LKTIBH A LKW irrigation in stating the laws for the journalists with the cooperation the present between/among /A Leo NIFIL/ liar A LKTIBH A LKW irrigation and the Lebanese army.	Praised leader A LKTIBH A LKW battalion for the journalists about cooperation present among /A NIFIL/ and A LKTIBH A LKW battalion and the Lebanese army.	Praised leader A LKTIBH A LKW battalion for the journalists about cooperation present among UNIFIL and A LKTIBH A LKW battalion and the Lebanese army.	In a press conference the head of the Korean battalion praised the cooperation between UNIFIL, the Korean battalion and the Lebanese army.
liar comes Ali anchored its/his list my journalistic/journalist the Lebanese NBIL the Moorish which issued upside-down gulf war my initial pull hand editor in Paris.	Comes highest in list the journalist the Lebanese NBIL the Moorish who issued after gulf war the first pull hand newspaper in Paris.	Comes highest in list the journalist the Lebanese Nabil al-Maghribi who issued after gulf war the first pull hand newspaper in Paris.	At the top of the list is the Lebanese journalist Nabil al-Maghribi, who issued after the First Gulf War Al-Muharrir newspaper from Paris.
attacked colonist liar N extreme who colonist H there went what IEIR went QOH south east country treads south A friend ba LDdFH inexperienced bey the today.	Attacked extreme settlers H there what IEIR located south east town south A friend ba LDdFH in bank.	Attacked extreme settlers Havat Yair located south east town south A friend ba LDdFH in bank.	Extremist settlers from the settlement of Havat Yair located in the southeast of Yata town south of Hebron in the West Bank attacked.

Table A-1. MT output, reference translations, control and enhanced stimulus sets (continued).

demanded spokesman name its/his movement continental Mohammad Yt the Korean with oppression Ali Kabul my avoidance danger murder BQIH A LMKhTtW yen.	Spokesman for movement continental Mohammad Yt ask the Korean to push Ali Kabul avoid danger to murder the remaining hostages.	Spokesman for movement Qari Mohammad Yousuf ask the Korean to push Ali Kabul avoid danger to murder the remaining hostages.	Spokesman for the movement Qadri Mohammad Youssef asked the Koreans to pressure Kabul to avoid the risk of killing the rest of the hostages.

Appendix B.

Appendix B includes the instructions to the subjects.

Thank you for taking the time to help improve machine translation.

You will be asked to read sentences that have been translated from Arabic to English using a machine. Each machine translation will be accompanied by a translation of the same Arabic sentence, but done by a certified bilingual human translator and is considered the translation "gold standard." So, each sentence in Arabic is translated by the human translator and by a machine translation system. You will see both of these translations. Your task is to judge how the machine translation compares to the human translation.

Of interest is the degree to which the machine translation conveys the meaning present in the human translation. The machine translation may not contain good, natural-sounding English like the human translation but you need to overlook that. The question to ask yourself is, "Do I get the same meaning from the machine translation as I do from the human translation?"

Let's look at some examples.

Human Translation:

Mr. Goldman visited his uncle Ralph on Tuesday in Paris.

Machine translation:

Tuesday, Mr. Gold in Paris to visit his uncle, Ralph.

In this example, most all of the meaning available in the human translation is also available in the machine translation. "Mr. Goldman" is incorrectly translated as "Mr. Gold." The human translation is in the past tense and the machine translation is in either the present or future tense. On balance, though, nearly all the meaning survives the machine translation. The readability of the machine translation is not great but, again, we want you to ignore that.

In brief, the pros and cons of this translation are:

Pros: "uncle Ralph," "Tuesday," and "Paris" are all correctly translated

Cons: "Mr. Goldman" is incorrectly translated as "Mr. Gold"

Let's look at another example.

Human translation:

When the 82nd Airborne jumped at Market Garden, General Gavin was the first one out of the plane.

Machine translation:

82 surge in the market when the Hanging Gardens, General Gavin is the first one out of the plane.

Here less information survives the machine translation. The fact that General Gavin jumped out of the plane first is in the machine translation even though the tense has been changed from past to present. However, "82nd Airborne" and "Market Garden" have been lost. A student of World War II history may be able to make sense of the machine translation but the reader should be able to understand the meaning of the translation without any special knowledge.

Pros: "General Gavin" is correctly translated; what General Gavin did is correctly translated

Cons: "82nd Airborne" and "Market Garden" are not correctly translated

Another example.

Human translation:

Major Hassan reported to Colonel Ali that a dozen Humvees located in Al Asad Base aren't ready.

Machine translation:

Transfer to Colonel Hassan leading to a dozen cars Alhmralamugodh base Assad not ready.

This machine translation gets many things wrong. The person "Hassan" survives the translation but "Colonel Ali" does not. The rank of Hassan is changed from Major to Colonel. It seems that 12 cars (that are actually Humvees) are being transferred to (now) Colonel Hassan—a meaning not in the human translation. We have no idea what Alhmralamugodh is. There is a reference to base Assad (a mistranslation of Al Asad) not being ready when, in truth, the vehicles aren't ready, not the base.

Pros: The name "Hassan" survives translation; 12 vehicles, of some description, are mentioned

Cons: Hassan's rank should be Major, not Colonel; "Colonel Ali" and "Humvees" are not translated; "Al Asad" is translated as "Assad" (similar but different); the machine translation refers to a "transfer" which is not mentioned in the human translation.

As you can see from these three actual examples, the amount of meaning retained in a machine translation can vary widely. So how are you to assign a value to each sample machine translation? The answer follows.

There is a final example on the next page (page 4) but finish reading this page before looking at it. As with the examples before, consider how much of the meaning present in the human translation is also present in the machine translation. Score this as a number between 0 (zero) and 10. You may use integers, fractions, or decimal numbers (for example, these numbers would

all be acceptable scores: .45, ¾, 3, 5½, 8.25). Please do not use negative numbers. A score of 0 would mean that in no sense does the machine translation provide even a hint of the meaning expressed in the human translation. Of course a score of 10 would reflect a machine translation that perfectly conveys all of the meaning in the human translation (neither adding nor losing any meaning).

Now turn to the next page, read the example, and write down the number you feel represents how much of the meaning in the human translation is contained in the machine translation.

Human translation:

General Mohamed led the Shwnies during February's River Blitz.

Machine translation:

Led by General Mohamed shwnies during a raid february's stream

Your score: _____

When finished, please turn to the next page.

Following are 20 machine translations with their associated human translations. Just as you determined a score for the last example, please write down the number you feel represents how much of the meaning in the human translation is contained in each machine translation.

Page from test booklet

Keeping in mind the score you gave to the example on page 4, what score would you give the following translation?

Human translation:

In addition to the Afghani forces there are also foreign forces that are headquartered in Ghazni.

Machine translation:

In addition to forces Afghan there are also foreign forces central station in Ghazni Province.

Your score: _____

Acronyms

EH	Expectancy Hypothesis
IH	Input Hypothesis
ILR	Interagency Language Roundtable
IPT	Integrated Process Teams
IV	independent variable
ME	Magnitude Estimation
METEOR	Metric for Evaluation of Translation with Explicit Ordering
MT	machine translation
NIST	National Institute of Standards and Technology

No. of Copies	Organization
1 PDF	ADMNSTR DEFNS TECHL INFO CTR ATTN DTIC OCP 8725 JOHN J KINGMAN RD STE 0944 FT BELVOIR VA 22060-6218
1	DARPA ATTN IXO S WELBY 3701 N FAIRFAX DR ARLINGTON VA 22203-1714
1 CD	OFC OF THE SECY OF DEFNS ATTN ODDRE (R&AT) THE PENTAGON WASHINGTON DC 20301-3080
1	US ARMY RSRCH DEV AND ENGRG CMND ARMAMENT RSRCH DEV AND ENGRG CTR ARMAMENT ENGRG AND TECHNLGY CTR ATTN AMSRD AAR AEF T J MATTS BLDG 305 ABERDEEN PROVING GROUND MD 21005-5001
1	US ARMY TRADOC BATTLE LAB INTEGRATION & TECHL DIRCTRT ATTN ATCD B 10 WHISTLER LANE FT MONROE VA 23651-5850
1	PM TIMS, PROFILER (MMS-P) AN/TMQ-52 ATTN B GRIFFIES BUILDING 563 FT MONMOUTH NJ 07703
1	US ARMY INFO SYS ENGRG CMND ATTN AMSEL IE TD F JENIA FT HUACHUCA AZ 85613-5300
1	COMMANDER US ARMY RDECOM ATTN AMSRD AMR W C MCCORKLE 5400 FOWLER RD REDSTONE ARSENAL AL 35898-5000
1	US GOVERNMENT PRINT OFF DEPOSITORY RECEIVING SECTION ATTN MAIL STOP IDAD J TATE 732 NORTH CAPITOL ST NW WASHINGTON DC 20402
1	US ARMY RSRCH LAB ATTN AMSRD ARL CI OK TP TECHL LIB T LANDFRIED BLDG 4600 ABERDEEN PROVING GROUND MD 21005-5066
1	DIRECTOR US ARMY RSRCH LAB ATTN AMSRD ARL RO EV W D BACH PO BOX 12211 RESEARCH TRIANGLE PARK NC 27709
7 HC 1 PDF	US ARMY RSRCH LAB ATTN AMSRD ARL CI IT J D WALRATH ATTN AMSRD ARL CI IT M VANNI (3 HC, 1 PDF) ATTN AMSRD ARL CI OK PE TECHL PUB ATTN AMSRD ARL CI OK TL TECHL LIB ATTN IMNE ALC IMS MAIL & RECORDS MGMT ADELPHI MD 20783-1197
1	DARPA J OLIVE 3701 N FAIRFAX DR ARLINGTON VA 22203-1714

Total: 20 (1 CD, 17 HCs, 2 PDFs)